# Data integration investigation

## Contents

# 1 Introduction

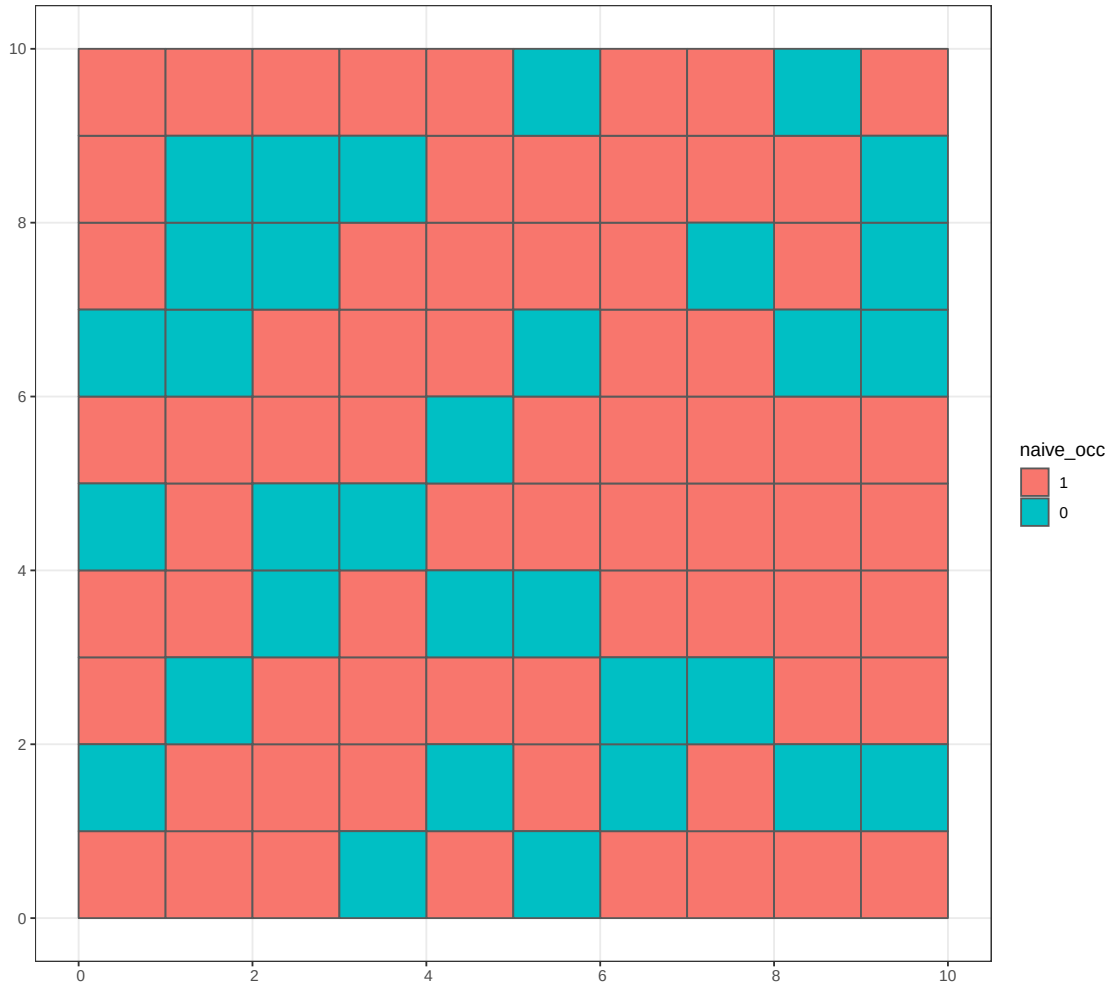
# 2 Traditional occupancy models

## 2.1 Two-stage occupancy

$$z_i \sim \text{Bernoulli}(\psi_i), \qquad \text{logit}(\psi_i) = x_i'\beta,$$

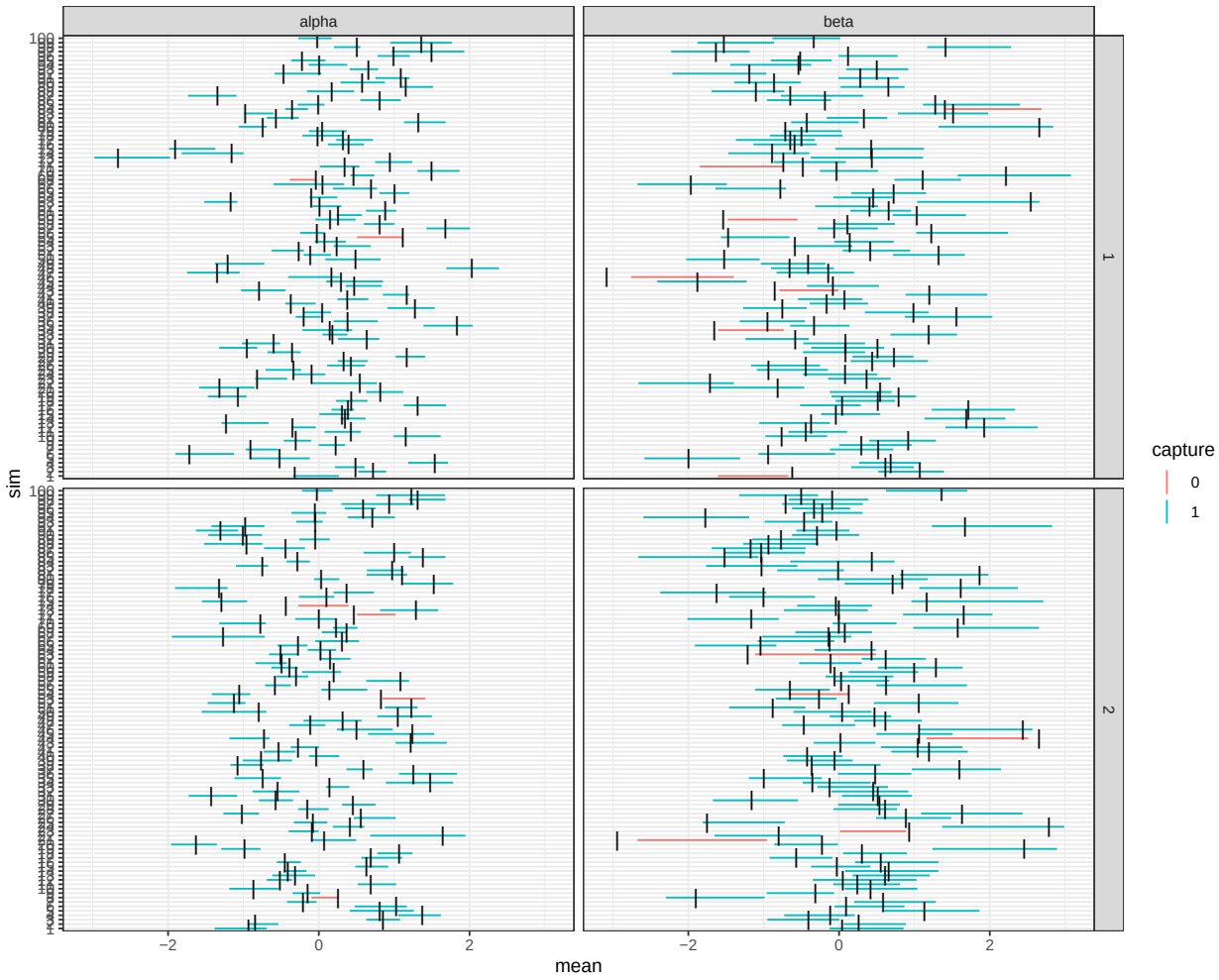
$$y_{ij} \sim \text{Bernoulli}(z_i p_{ij}), \qquad \text{logit}(p_{ij}) = w_{ij}'\alpha$$

### 2.1.1 Simulated data



### 2.1.2 Gibbs sampler simulation

Everything seems to be working.



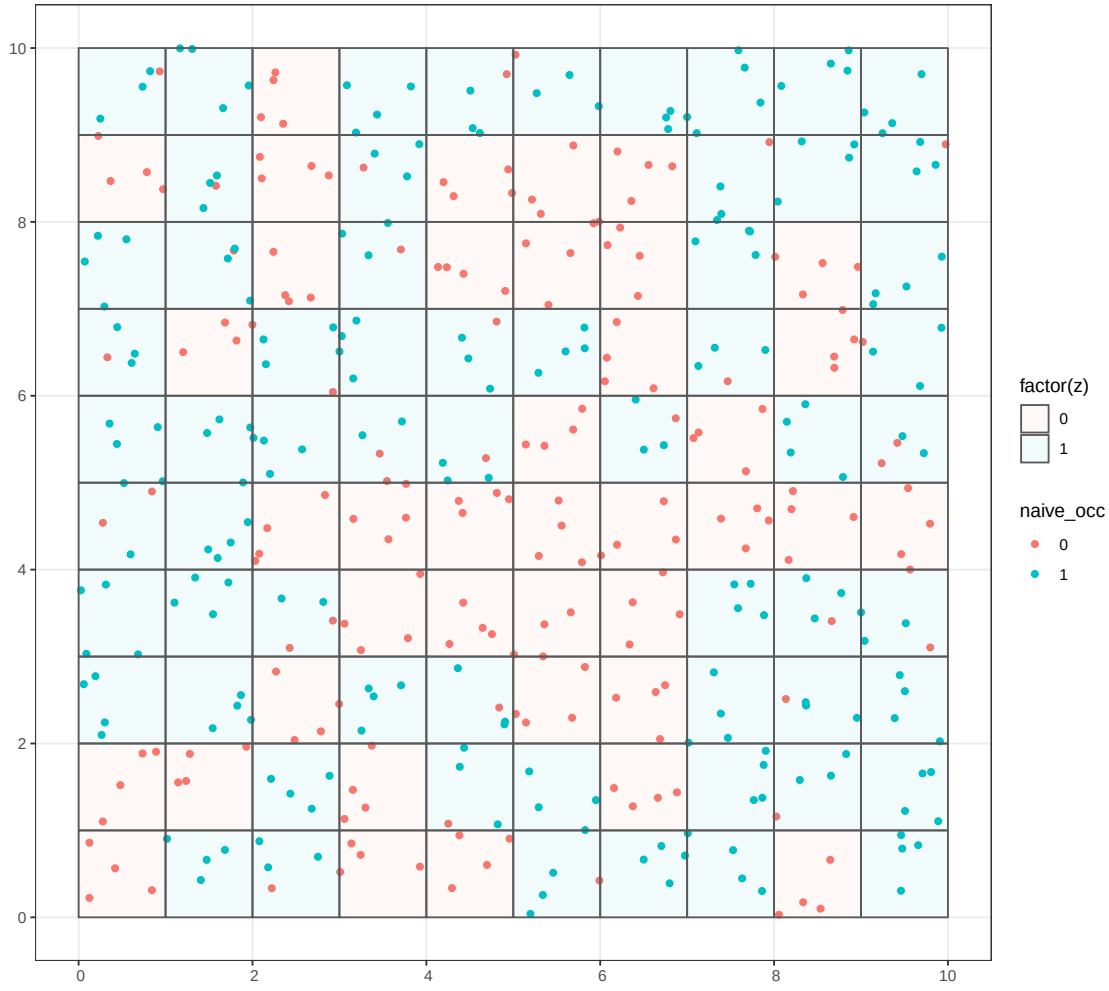

## 2.2 Three-stage occupancy

$$z_i \sim \text{Bernoulli}(\psi_i), \qquad \text{logit}(\psi_i) = x_i'\beta,$$

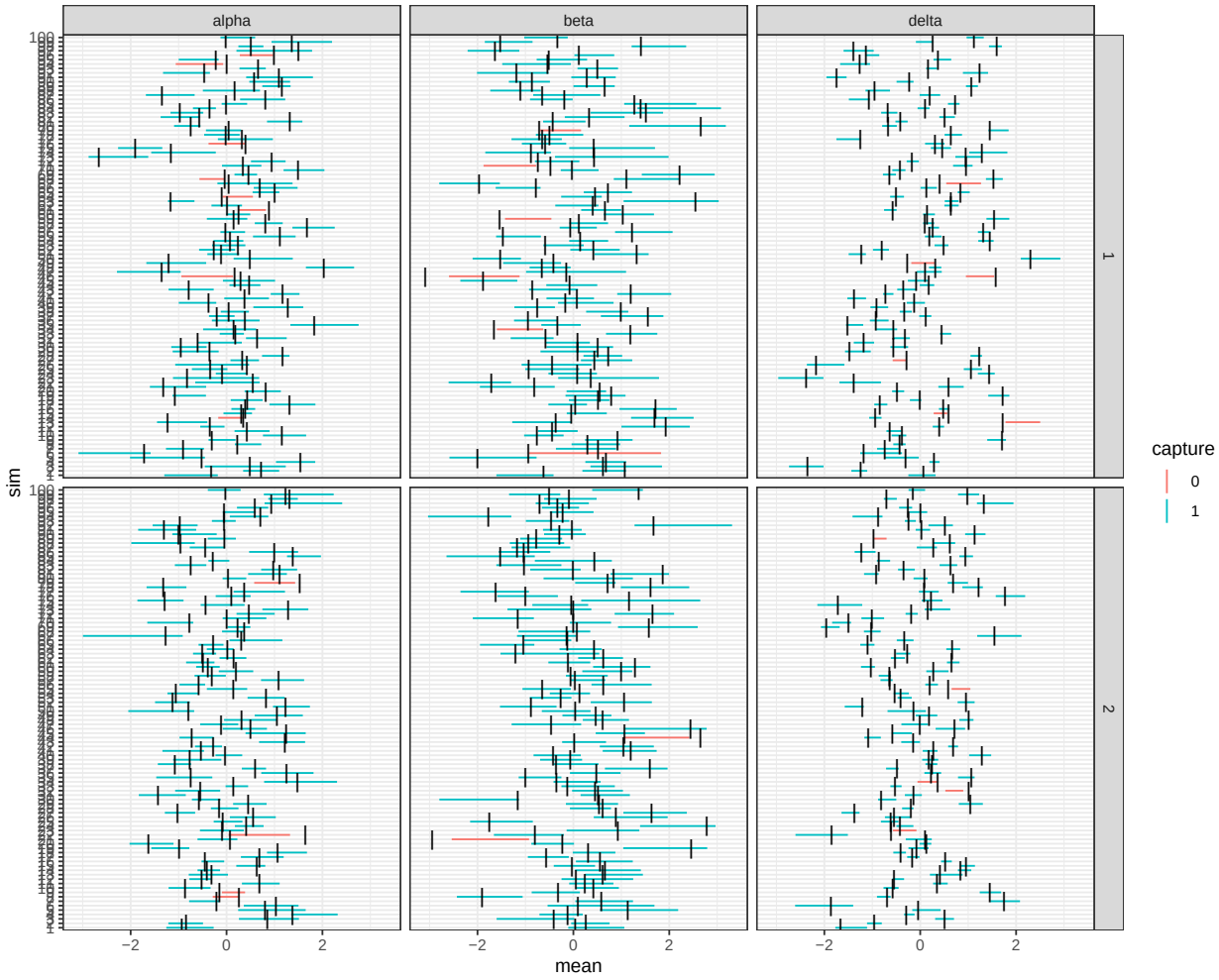
$$a_{ij} \sim \text{Bernoulli}(z_i\theta_{ij}), \qquad \text{logit}(\theta_{ij}) = w_{ij}'\alpha$$

$$y_{ijk} \sim \text{Bernoulli}(a_{ij}p_{ijk}), \qquad \text{logit}(p_{ijk}) = v_{ijk}'\delta$$

### 2.2.1 Simulated data



### 2.2.2 Gibbs sampler simulation




# 3 Count occupancy models

## 3.1 Three-stage count occupancy

$$z_i \sim \text{Bernoulli}(\psi_i), \qquad \text{logit}(\psi_i) = x_i'\beta,$$

$$a_{ij} \sim \text{Bernoulli}(z_i\theta_{ij}), \qquad \text{logit}(\theta_{ij}) = w_{ij}'\alpha$$

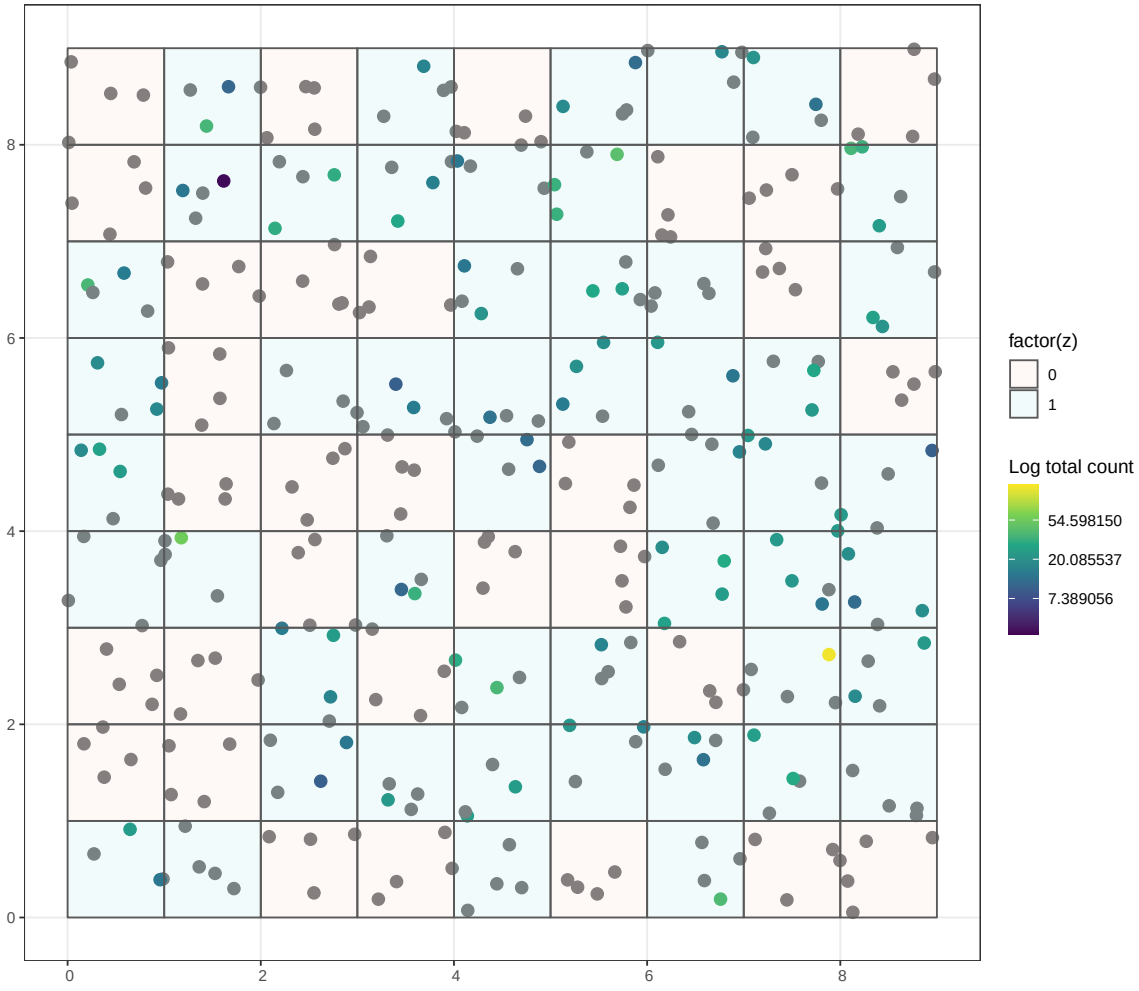
$$y_{ijk} \sim \text{NegBin}(r, a_{ij}p_{ijk}), \qquad \text{logit}(p_{ijk}) = v_{ijk}'\delta$$

where the NegBin$(r, p)$ distribution has density

$$p(y|r,p) = \binom{y+r-1}{y} p^r (1-p)^y$$
$$\propto p^r (1-p)^y$$

Note that assuming $\text{logit}(p_i) = x_i\beta$, $p_i^r(1-p_i)^{y_i} = \frac{[\exp(x_i\beta)]^r}{[1+\exp(x_i\beta)]^{y_i+r}}$.

### 3.1.1 Simulated data



### 3.1.2 Gibbs sampler - skip this section

**3.1.2.1 Derivations** Below, we derive the Gibbs step for regression coefficients given a negative binomial sampling model. Throughout, we omit subscripts to ease notation. Suppose $y \sim \text{NegBin}(r, p)$, where

$$\begin{aligned} p(y|r,p) &= \binom{y+r-1}{y} p^r (1-p)^y \\ &\propto p^r (1-p)^y, \end{aligned}$$

and assume a logit link function $p = \frac{\exp(x\beta)}{1+\exp(x\beta)}$. Theorem 1 from Polson, Scott and Windle (2013) states that if $\omega \sim PG(b, 0)$, then for all $a \in \mathbb{R}$,

$$\frac{(e^\psi)^a}{(1+e^\psi)^b} = 2^{-b} e^{\kappa\psi} \int_0^\infty e^{-\omega\psi^2/2} p(\omega) d\omega,$$

where $\kappa = a - b/2$.

We leverage Theorem 1 to construct a Gibbs sampler for negative binomial sampling models by introducing Polya-gamma distributed auxiliary variables, resulting in a conditionally Gaussian kernel for which Gibbs sampling techniques are well defined. First, we manipulate the likelihood.

$$\begin{aligned} p(y|r,p) &= \binom{y+r-1}{y} p^r (1-p)^y \\ &\propto p^r (1-p)^y \\ &= \left(\frac{\exp(x\beta)}{1+\exp(x\beta)}\right)^r \left(1 - \frac{\exp(x\beta)}{1+\exp(x\beta)}\right)^y \\ &= \left(\frac{\exp(x\beta)}{1+\exp(x\beta)}\right)^r \left(\frac{1}{1+\exp(x\beta)}\right)^y \\ &= \frac{(\exp(x\beta))^r}{(1+\exp(x\beta))^{r+y}} \end{aligned}$$

By Theorem 1 from Polson, Scott and Windle (2013),

$$\begin{aligned} \frac{(\exp(x\beta))^r}{(1+\exp(x\beta))^{r+y}} &= 2^{-(r+y)} e^{\kappa x\beta} \int_0^\infty e^{-\omega(x\beta)^2/2} p(\omega) d\omega \\ &\propto e^{\kappa x\beta} \int_0^\infty e^{-\omega(x\beta)^2/2} p(\omega) d\omega \end{aligned}$$

where $\kappa = r - (y+r)/2$. Now, consider the likelihood contribution for a single observation for the regression

coefficients $\beta$.

$$\begin{aligned} L_(\beta|y_i) &\propto \frac{(\exp(x_i\beta))^r}{(1+\exp(x_i\beta))^{r+y_i}} = 2^{-(r+y_i)} e^{\kappa_i x_i \beta} \int_0^\infty e^{-\omega_i (x_i\beta)^2/2} p(\omega_i) d\omega_i \\ &\propto e^{\kappa_i x_i \beta} \int_0^\infty e^{-\omega_i (x_i\beta)^2/2} p(\omega_i) d\omega_i \end{aligned}$$

where $\kappa_i = r - (y_i + r)/2$. To obtain our Gibbs step for the regression coefficients, we condition on the Polya-gamma auxiliary variables and consider $n$ observations.

$$\begin{aligned} p(\beta|\omega, y) &\propto p(\beta) \prod_{i=1}^n L_i(\beta|\omega_i, y_i) \\ &= p(\beta) \prod_{i=1}^n \exp\left\{\kappa_i x_i \beta - \omega_i (x_i\beta)^2/2\right\} \end{aligned}$$

Refresher on completing the square:

$$\begin{aligned} \exp\left\{\kappa_i x_i \beta - \omega_i \frac{(x_i\beta)^2}{2}\right\} &= \exp\left\{-\frac{\omega_i}{2}\left((x_i\beta)^2 - \frac{2\kappa_i}{\omega_i}(x_i\beta)\right)\right\} \\ &= \exp\left\{-\frac{\omega_i}{2}\left((x_i\beta)^2 - \frac{2\kappa_i}{\omega_i}(x_i\beta) + \frac{\kappa_i^2}{\omega_i^2} - \frac{\kappa_i^2}{\omega_i^2}\right)\right\} \\ &= \exp\left\{-\frac{\omega_i}{2}\left(\left(x_i\beta - \frac{\kappa_i}{\omega_i}\right)^2 - \frac{\kappa_i^2}{\omega_i^2}\right)\right\} \\ &\propto \exp\left\{-\frac{\omega_i}{2}\left(x_i\beta - \frac{\kappa_i}{\omega_i}\right)^2\right\} \end{aligned}$$

Returning to the full conditional distribution of $\beta$

$$\begin{aligned} p(\beta|\omega, y) &\propto p(\beta) \prod_{i=1}^n L_i(\beta|\omega_i, y_i) \\ &= p(\beta) \prod_{i=1}^n \exp\left\{\kappa_i x_i \beta - \omega_i (x_i\beta)^2/2\right\} \\ &\propto p(\beta) \prod_{i=1}^n \exp\left\{-\frac{\omega_i}{2}\left(x_i\beta - \frac{\kappa_i}{\omega_i}\right)^2\right\} \end{aligned}$$

Let $z_i = \frac{\kappa_i}{\omega_i}$. Then $p(\beta|\omega, y) \propto p(\beta) \prod_{i=1}^n \exp\left\{-\frac{\omega_i}{2}\left(z_i - x_i\beta\right)^2\right\}$, which is the kernel of a $N(x_i\beta, \frac{1}{\omega_i})$ distribution. Combining all $n$ observations into matrix notation, we have

$$p(\beta|\omega, y) \propto p(\beta) \exp\left\{-\frac{1}{2}(z - X\beta)'\Omega(z - X\beta)\right\}$$

where $z = \left(\frac{1}{2\omega_1}(r - y_1), ..., \frac{1}{2\omega_n}(r - y_n)\right)$ and $\Omega = \text{diag}(\omega_1, ..., \omega_n)$.

This same process is replicated at the first two levels of the occupancy hierarchy, but with binomial sampling models. See Polson, Scott and Windle (2013) for full details. Additionally, full conditional distributions are required for the Polya-gamma auxiliary variables and the dispersion parameter for the negative binomial distribution. The full conditional distribution for the Polya-gamma auxiliary variables is

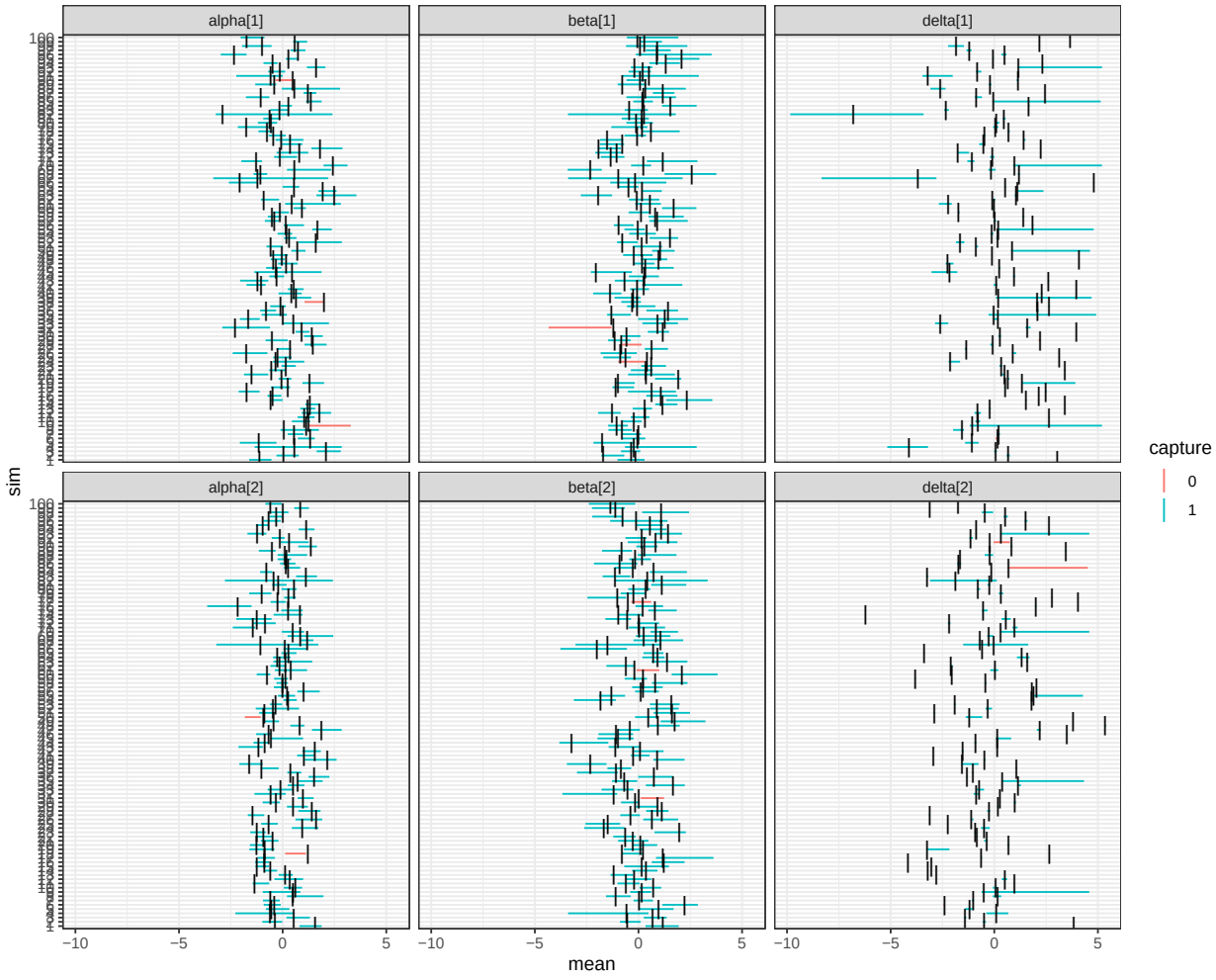
$$\omega_i|\beta \sim PG(y_i + r, x_i\beta),$$

see Polson, Scott and Windle (2013) for full detail. To sample the dispersion parameter, we implement the method described by Zhou et al. (2012).

### 3.1.3 Probabilistic programming language

Since we are using a PPL to sample the posterior distribution, conditional conjugacy between the priors and sampling model is not required. Therefore, we replace the bottom level of the hierarchy with a Poisson sampling model for the counts.

$$\begin{aligned} z_i &\sim \text{Bernoulli}(\psi_i), \qquad \text{logit}(\psi_i) = x_i'\beta, \\ a_{ij} &\sim \text{Bernoulli}(z_i\theta_{ij}), \qquad \text{logit}(\theta_{ij}) = w_{ij}'\alpha \\ y_{ijk} &\sim \text{Poisson}(a_{ij}\lambda_{ijk}), \qquad \log(\lambda_{ijk}) = v_{ijk}'\delta \end{aligned}$$




# 4 Integrated count occupancy

## 4.1 Continuous disease prevalence surface

We extend the three-stage count occupancy model to accommodate point-referenced disease surveillance data.

$$z_i \sim \text{Bernoulli}(\psi_i), \qquad \text{logit}(\psi_i) = x_i'\beta,$$

$$a_{ij} \sim \text{Bernoulli}(z_i\theta_{ij}), \qquad \text{logit}(\theta_{ij}) = w_{ij}'\alpha$$

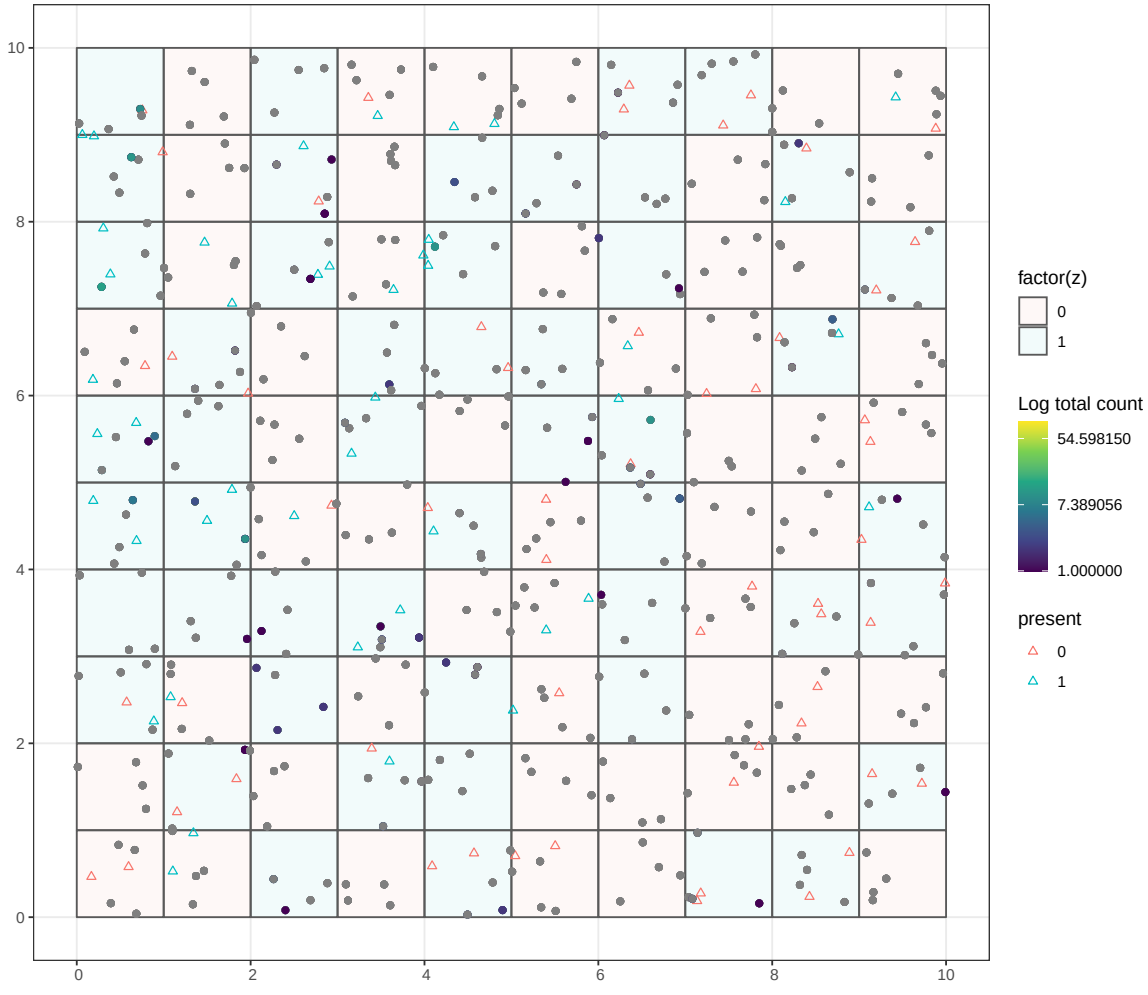
$$y_{ijk} \sim \text{Poisson}(a_{ij}\lambda_{ijk}), \qquad \log(\lambda_{ijk}) = v_{ijk}'\delta + \eta_{ij}\gamma$$

where $p_{ij}$ represents the disease prevalence and

$$\text{logit}(p_{ij}) = \eta_{ij}$$

$$\eta \sim \mathcal{N}(0, \Sigma)$$

where $\Sigma_{ij} = \sigma^2 \exp\left\{-\frac{d_{ij}}{\phi}\right\}$ and $d_{ij}$ represents the distance between sample locations $i$ and $j$.



# References

Polson, N.G., Scott, J.G. and Windle, J. (2013) Bayesian inference for logistic models using Pólya–gamma latent variables. *Journal of the American Statistical Association*, **108**, 1339–1349.

Zhou, M., Li, L., Dunson, D. and Cain, L. (2012) Lognormal and gamma mixed negative binomial regression. *Proc Int Conf Mach Learn*, **2012**, 1343–1350.